

# TESTING FOR HOMOGENEITY IN MIXTURE MODELS

JIAYING GU, ROGER KOENKER, AND STANISLAV VOLGUSHEV

**ABSTRACT.** Statistical models of unobserved heterogeneity are typically formalized as mixtures of simple parametric models and interest naturally focuses on testing for homogeneity versus general mixture alternatives. Many tests of this type can be interpreted as  $C(\alpha)$  tests, as in Neyman (1959), and shown to be locally, asymptotically optimal. A unified approach to analysing the asymptotic behavior of such tests will be described, employing a variant of the LeCam LAN framework. These  $C(\alpha)$  tests will be contrasted with a new approach to likelihood ratio testing for mixture models. The latter tests are based on estimation of general (nonparametric) mixture models using the Kiefer and Wolfowitz (1956) maximum likelihood method. Recent developments in convex optimization are shown to dramatically improve upon earlier EM methods for computation of these estimators, and new results on the large sample behavior of likelihood ratios involving such estimators yield a tractable form of asymptotic inference. We compare performance of the two approaches identifying circumstances in which each is preferred.

## 1. INTRODUCTION

Given a simple parametric density model,  $f(x, \vartheta)$ , for iid observations,  $X_1, \dots, X_n$ , there is a natural temptation to complicate the model by allowing the parameter,  $\vartheta$ , to vary with  $i$ . In the absence of other, e.g. covariate, information that would distinguish the observations from one another it may be justifiable to view the  $\vartheta$ 's as drawn at random. Inference for such mixture models is complicated by a variety of problems, notably their lack of identifiability. Two dominant approaches exist: Neyman's  $C(\alpha)$  and the likelihood ratio test.  $C(\alpha)$  is particularly attractive for testing homogeneity against general forms of heterogeneity for the parameter  $\vartheta$ , such tests have a relatively simple asymptotic theory, and are generally easy to compute. The LRT, in contrast, is more easily adapted to compound null hypotheses, but has a much more complicated limiting behavior, and is generally more difficult to compute.

We will argue that recent developments in convex optimization have dramatically reduced the computational burden of the LRT approach for general, nonparametric alternatives. We will present a tractable large-sample theory for the LRT that conforms well to our simulation evidence, and exhibits both good size and power performance. Comparisons throughout with  $C(\alpha)$ , which is asymptotically locally optimal, demonstrate that the LRT can be a highly effective complementary approach.

---

Version: February 8, 2013. This research was partially supported by NSF grant SES-11-53548. This research was conducted while the third author was a visiting scholar at UIUC. He is very grateful to the Statistics and Economics departments for their hospitality.

## 2. LIKELIHOOD RATIO TESTS FOR MIXTURE MODELS

Lindsay (1995) offers a comprehensive overview of the vast literature on mixture models. He traces the idea of maximum likelihood estimation of a *nonparametric* mixing distribution  $F$ , given random samples from the mixture density,

$$(1) \quad g(x) = \int \varphi(x, \vartheta) dF(\vartheta),$$

to Robbins (1950). Kiefer and Wolfowitz (1956) filled in many details of the Robbins proposal and yet only with Laird (1978) did a viable computational strategy emerge for it. The EM method proposed by Laird has been employed extensively in subsequent work, e.g. Heckman and Singer (1984) and Jiang and Zhang (2009), even though it has been widely criticized for its slow convergence. Recently Koenker and Mizera (2012) have noted that the Kiefer-Wolfowitz estimator can be formulated as a convex optimization problem and solved very efficiently by interior point methods. Recent work by Liu and Shao (2003) and Azaïs, Gassiat, and Mercadier (2009) has clarified the limiting behavior of the LRT for general classes of alternatives, and taken together these developments offer a fresh opportunity to explore the LRT for inference on mixtures.

It seems ironic that many of the difficulties inherent in maximum likelihood estimation of finite parameter mixture models vanish when we consider nonparametric mixtures. The notorious multimodality of parametric likelihood surfaces is replaced by a much simpler, strictly convex optimization problem possessing a unique, unimodal solution. It is of obvious concern that consideration of such a wide class of alternatives may depress the power of associated tests; we will see that while there is some loss of power when compared to more restricted parametric LRTs, the loss is typically modest, a small price to pay for power against a broader class of alternatives. We will see that by comparison with  $C(\alpha)$  tests that are also designed to detect general alternatives, the LRT can be competitive.

**2.1. Maximum Likelihood Estimation of General Mixtures.** Suppose that we have iid observations,  $X_1, \dots, X_n$  from the mixture density (1), the Kiefer-Wolfowitz MLE requires us to solve,

$$\min_{F \in \mathcal{F}} \left\{ - \sum_{i=1}^n \log g(x_i) \right\},$$

where  $\mathcal{F}$  is the (convex) set of all mixing distributions. The problem is one of minimizing the sum of convex functions subject to linear equality and inequality constraints. The dual to this (primal) convex program proves to be somewhat more tractable from a computational viewpoint, and takes the form,

$$\max_{\nu \in \mathbb{R}^n} \left\{ \sum_{i=1}^n \log \nu_i \mid \sum_{i=1}^n \nu_i \varphi(x_i, \vartheta) \leq n, \quad \text{for all } \vartheta \right\}$$

See Lindsay (1983) and Koenker and Mizera (2012) for further details. This variational form of the problem may still seem rather abstract since it appears that we need to check an infinite number of values of  $\vartheta$ , for each choice of the vector,  $\nu$ . However, it suffices in

applications to consider a fine grid of values  $\{\vartheta_1, \dots, \vartheta_m\}$  and write the primal problem as

$$\min_{f \in \mathbb{R}^m, g \in \mathbb{R}^n} \left\{ - \sum_{i=1}^n \log(g_i) \mid Af = g, f \in \mathcal{S} \right\}$$

where  $A$  is an  $n$  by  $m$  matrix with elements  $\varphi(x_i, \vartheta_j)$  and  $\mathcal{S} = \{s \in \mathbb{R}^m \mid 1^\top s = 1, s \geq 0\}$  is the unit simplex. Thus,  $\hat{f}_j$  denotes the estimated mixing density evaluated at the grid point,  $\vartheta_j$  and  $\hat{g}_i$  denotes the estimated mixture density evaluated at  $x_i$ . The dual problem in this discrete formulation becomes,

$$\max_{\nu \in \mathbb{R}^n} \left\{ \sum_{i=1}^n \log \nu_i \mid A^\top \nu \leq n1_m, \nu \geq 0 \right\}.$$

Primal and dual solutions are immediately recoverable from the solution to either problem. Interior point methods such as those provided by PDCO of Saunders (2003) and Mosek of Andersen (2010), are capable of solving dual formulations of typical problems with  $n = 200$  and  $m = 300$  in less than one second.<sup>1</sup>

Solutions to the nonparametric MLE problem of Kiefer and Wolfowitz produce estimates of the mixing distribution,  $F$ , that are discrete and possessing only a few mass points. A theoretical upper bound on the number of these atoms of  $n$  was established already by Lindsay (1983), but in practice the number is actually observed to be far fewer. It may seem surprising, perhaps even disturbing, that even when the true mixing distribution has a smooth density, our estimates of that density is discrete with only a handful of atoms. This may appear less worrying if we consider a more explicit example. Suppose that we have a location mixture of Gaussians,

$$g(x) = \int \phi(x - \mu) dF(\mu),$$

so we are firmly in the deconvolution business, a harsh environment notorious for its poor convergence rates. One interpretation of this is that good approximations of the mixture density  $g$  can be achieved by relatively simple discrete mixtures with only a few atoms. For many applications estimation of  $g$  is known to be sufficient: this is quite explicit for example for empirical Bayes compound decision problems where Bayes rules depend entirely on the estimated  $\hat{g}$ . Of course given our discrete formulation of the Kiefer-Wolfowitz problem, we can only identify the location of atoms up to the scale of the grid spacing, but we believe that the  $m \approx 300$  grid points we have been using are probably adequate for most applications.

Given a reliable maximum likelihood estimator for the general nonparametric mixture model it is of obvious interest to know whether an effective likelihood ratio testing strategy can be developed. This question has received considerable prior attention, again Lindsay (1995) provides an authoritative overview of this literature. However, more recently work by Liu and Shao (2003) and Azaïs, Gassiat, and Mercadier (2009) have revealed new features of the asymptotic behavior of the likelihood ratio for mixture settings that enable one to derive asymptotic critical values for the LRT.

---

<sup>1</sup> The R empirical Bayes package **REBayes**, Koenker (2012), is available from the second author on request. It is based on the **RMosek** package of Friberg (2012), and was used for all of the computations reported below.

**2.2. Asymptotic Theory of Likelihood Ratios for General Mixtures.** Consider the general problem of testing that observed data come from a family of distributions  $(F_\vartheta)_{\vartheta \in \Theta_1}$  against the alternative that the distribution is  $F_\vartheta$  for some  $\vartheta \in \Theta_2 \setminus \Theta_1$  where we assume  $\Theta_1 \subset \Theta_2$ . Liu and Shao (2003) provide tools that allow one to derive the limiting distribution of the likelihood ratio test statistic under very general conditions. In particular,  $\Theta_1, \Theta_2$  need not be subsets of  $\mathbb{R}^d$  but are allowed to be subsets of general metric spaces.

Denote the “true” density of the data by  $f$  and start by considering a point null hypothesis, i.e.  $H_0 : f = f_0$  for some density  $f_0$  (note that we do not assume that there exists a unique parameter  $\vartheta$  corresponding to  $f_\vartheta = f_0$ , although in the setting considered here this will in fact turn out to be the case). Denote by  $\ell_\vartheta(x) := f_\vartheta(x)/f_0(x)$  the likelihood ratio and consider the test statistic for  $H_0$  against  $H_1 : f = f_\vartheta \neq f_0$

$$\sup_{\vartheta \in \Theta} L_n(\vartheta), \quad L_n(\vartheta) := \sum_{i=1}^n \log \ell_\vartheta(X_i).$$

Adapting Theorem 3.1 of Liu and Shao (2003) yields the following theorem.

**Theorem 2.1.** *Define the classes of functions*

$$\mathcal{F}_{\Theta, \varepsilon} := \left\{ S_\vartheta := \frac{\ell_\vartheta - 1}{D(\vartheta)} \mid 0 < D(\vartheta) \leq \varepsilon \right\}$$

and

$$\mathcal{F}_{\Theta, 0} := \left\{ S \in L^2 : \exists (\vartheta_n)_{n \in \mathbb{N}} \subset \Theta \text{ s.t. } D(\vartheta_n) = o(1), \left\| \frac{\ell_{\vartheta_n} - 1}{D(\vartheta_n)} - S \right\|_2 = o(1) \right\}$$

with  $D^2(\vartheta) := \mathbb{E}[(\ell_\vartheta - 1)^2]$ . Assume that the following three conditions hold

- (1) For sufficiently small  $\varepsilon > 0$ , the class  $\mathcal{F}_{\Theta, \varepsilon}$  is Donsker.
- (2) For any sequence  $(\vartheta_n)_{n \in \mathbb{N}} \subset \Theta$  with  $D(\vartheta_n) = o(1)$  there exists a subsequence  $(\vartheta_{n_k})_{k \in \mathbb{N}} \subset \Theta$  and a function  $S \in \mathcal{F}_{\Theta, 0}$  with  $\|S_{\vartheta_{n_k}} - S\|_2 = o(1)$ .
- (3) For any  $S \in \mathcal{F}_{\Theta, 0}$  there exists a path  $\{\vartheta(t, S) : 0 < t \leq \varepsilon\} \subset \Theta$  such that  $t \mapsto \ell_{\vartheta(t, S)}$  is continuous, with respect to the  $L^2$  norm,  $D(\vartheta(t, S)) > 0$  for all  $t > 0$  and  $\lim_{t \rightarrow 0} S_{\vartheta(t, S)} = S$  in  $L^2$ .

Then

$$2 \sup_{\vartheta \in \Theta} L_n(\vartheta) \rightsquigarrow \left( \sup_{S \in \mathcal{F}_{\Theta, 0}} W_S \vee 0 \right)^2$$

where  $(W_S)_{S \in \mathcal{F}_{\Theta, 0}}$  denotes a centered Gaussian process with covariance structure  $\text{Cov}(W_f, W_g) = \mathbb{E}[f(X)g(X)]$ .

**Remark 2.2.** The condition (2) is called *completeness* by Liu and Shao. The condition (3) is slightly different from the assumption called *continuous sample paths* in Definition 2.4 of Liu and Shao (2003). However, a closer look at the relevant proofs reveals that condition (3) is in fact sufficient.

The asymptotic distribution of likelihood ratio tests with composite null hypothesis, i.e. the test  $\vartheta \in \Theta_2$  versus  $\vartheta \in \Theta_1$  is then given by,

$$\left( \sup_{S \in \mathcal{F}_{\Theta_2}} W_S \vee 0 \right)^2 - \left( \sup_{S \in \mathcal{F}_{\Theta_1}} W_S \vee 0 \right)^2.$$

The two main challenges in applying this result thus consist in describing the classes  $\mathcal{F}_{\Theta}$  and in verifying the technical assumptions. Let us start by considering a classical example where the limiting distribution is known.

**Example 2.3.** Assume that all parameters are  $\mathbb{R}^d$ -valued and identifiable, i.e.  $\vartheta_1 = \vartheta_2$  if and only if  $f_{\vartheta_1} = f_{\vartheta_2}$ , and that  $D(\vartheta_n) = o(1)$  if and only if  $\vartheta_n \rightarrow \vartheta_0$ . Moreover, let  $f_{\vartheta}$  be “nice” (sufficiently smooth with respect to  $\vartheta$  in a uniform sense). For simplicity, assume that the true parameter is zero and that it is an interior point of  $\Theta$ . A Taylor expansion then yields (the first equality holding in an  $L^2$ -sense)

$$\ell_{\vartheta_n} - 1 = \vartheta_n^\top \ell'_0 + o(\|\vartheta_n\|), \quad D^2(\vartheta_n) = \vartheta_n^\top \mathbb{E}[\ell'_0(\ell'_0)^\top] \vartheta_n + o(\|\vartheta_n\|^2),$$

and the class of functions  $\mathcal{F}_{\Theta}$  is identified as

$$\mathcal{F}_{\Theta} = \left\{ \frac{v^\top \ell'_0}{(v^\top \mathbb{E}[\ell'_0(\ell'_0)^\top] v)^{1/2}} \mid v \in S^{d-1} \right\}.$$

Identify this class of functions with the sphere  $S^{d-1}$ . The covariance structure of the Gaussian process  $W$  is now given by

$$\text{Cov}(W_v, W_w) = \frac{v^\top \mathbb{E}[\ell'_0(\ell'_0)^\top] w}{(v^\top \mathbb{E}[\ell'_0(\ell'_0)^\top] v)^{1/2} (w^\top \mathbb{E}[\ell'_0(\ell'_0)^\top] w)^{1/2}}.$$

The seemingly formidable “Gaussian process” thus turns out to be merely a collection of scaled linear combinations of a  $d$ -dimensional normal distribution. More precisely, its distribution coincides with that of  $(Z_v)_{v \in S^{d-1}}$  where

$$Z_v := \frac{v^\top Y}{\text{Var}(v^\top Y)^{1/2}}, \quad Y \sim \mathcal{N}(0, \mathbb{E}[\ell'_0(\ell'_0)^\top]).$$

Finally, writing  $Y = \mathbb{E}[\ell'_0(\ell'_0)^\top]^{1/2} \tilde{Y}$  with  $\tilde{Y} \sim \mathcal{N}(0, I_d)$  yields the expected  $\chi_d^2$  distribution

$$\left( \sup_{S \in \mathcal{F}_{\Theta}} W_S \vee 0 \right)^2 \sim \tilde{Y}_1^2 + \dots + \tilde{Y}_d^2 \sim \chi_d^2.$$

Let us now turn to the more specialized case of testing homogeneity against arbitrary mixtures, i.e.

$$H_0 : f = f_\mu \text{ for some } \mu \in M \quad \text{against} \quad H_1 : f(x) = \int_M f_\mu(x) dG(\mu), \quad G \in P_M \setminus P_M^{(1)}$$

where  $P_M$  denotes the set of probability measures on  $M$ , and  $P_M^{(k)}$  the set of distribution functions that have exactly  $k$  mass points. Under  $H_0$ , there exists a  $\mu_0 \in M$  with  $f = f_{\mu_0}$ . The parameter set  $\Theta$  can now be identified with the set of measures  $P_M$ , that are identified with their distribution functions. The symbols  $f_G, \ell_G$  will be used to denote the mixture density and likelihood ratio corresponding to the distribution  $G$ . Abusing notation, we will

use the symbol  $\ell_\mu$  to denote both, the function  $x \mapsto f_\mu(x)/f_{\mu_0}(x)$  and the quantity  $\ell_{\delta_\mu}$ , with  $\delta_\mu$  denoting the distribution with point mass at  $\mu$ .

In principle, the results of Liu and Shao (2003) can be applied in this situation under rather general conditions on the mixture and mixing distributions. A closely related approach was recently taken by Azaïs, Gassiat, and Mercadier (2009), who derive the distribution of the likelihood ratio test for a single distribution against arbitrary mixtures under fairly general conditions.

**Example 2.4.** Consider mixtures of  $\mathcal{N}(\mu, 1)$  distributions and assume that  $M = [L, U]$  with  $0 \in M$ . According to theorem 3 in Azaïs, Gassiat, and Mercadier (2009), the asymptotic distribution of the log-likelihood ratio test statistic

$$2 \left( \sup_{G \in P_M} \sum_{i=1}^n \log \ell_G(X_i) - \sup_{G \in P_M^{(1)}} \sum_{i=1}^n \log \ell_G(X_i) \right)$$

under the null of  $X_i \sim \mathcal{N}(0, 1)$  i.i.d. is given by

$$D = \left( \sup_{G \in P_M} (V_G)_+ \right)^2 - Y_1^2$$

where  $(V_G)_{G \in P_M}$  is the Gaussian process given by

$$V_G := \left( \sum_{k=1}^{\infty} \frac{Y_k \kappa_k(G)}{(k!)^{1/2}} \right) / \left( \sum_{k=1}^{\infty} \frac{\kappa_k^2(G)}{k!} \right)^{1/2}$$

with  $Y_1, Y_2, \dots$  denoting i.i.d.  $\mathcal{N}(0, 1)$  distributed random variables,  $\kappa_k(G) := \int_M \mu^k dG(\mu)$  and  $x_+$  denoting the positive part of  $x$ . As we will show in the appendix, there is a simpler expression for the distribution of  $D$ . More precisely, we will demonstrate that

$$(2) \quad D \stackrel{\mathcal{D}}{=} \sup_{G \in P_M} \left( \left( \left( \sum_{k=2}^{\infty} \frac{Y_k \kappa_k(G)}{(k!)^{1/2}} \right)_+ \right)^2 / \sum_{k=2}^{\infty} \frac{\kappa_k^2(G)}{k!} \right).$$

Approximating the distribution function  $G$  on  $M$  by a discrete distribution function with masses  $p_1, \dots, p_N$  on a fine grid  $m_1, \dots, m_N$  leads to the approximation

$$D \approx \sup_{p_1, \dots, p_N} \left( \left( \left( \sum_{j=1}^N p_j \sum_{k=2}^{\infty} \frac{Y_k m_j^k}{(k!)^{1/2}} \right)_+ \right)^2 / \sum_{i,j=1}^N p_i p_j \sum_{k=2}^{\infty} \frac{(m_j m_i)^k}{k!} \right).$$

In particular, maximizing the right-hand side with respect to  $p_1, \dots, p_N$  under the constraints  $p_i \geq 0, \sum p_i = 1$  for fixed grid  $m_1, \dots, m_N$  can be formulated as a quadratic optimization problem of the form

$$\min_p p^\top A p \quad \text{under} \quad p_i \geq 0, \quad p^\top b = 1$$

where  $p = (p_1, \dots, p_N)$ ,  $A_{ij} = \sum_{k=2}^{\infty} \frac{(m_j m_i)^k}{k!}$ ,  $b_i = \sum_{k=2}^{\infty} \frac{Y_k m_i^k}{(k!)^{1/2}}$ , if  $\max_i b_i > 0$ . If  $\max_i b_i \leq 0$ , we can set  $D = 0$ . This suggests a practical way of simulating critical values after replacing the infinite sum by a finite approximation and avoiding the grid point 0. The table below contains simulated critical values in some particular settings. All results are based on 10,000 simulation runs with the sums for  $A$  and  $b$  cut off at  $k = 25$  and grids with 200 points equally spaced points excluding the point 0.

$M$	90%	95%	99%
$[-1,1]$	2.75	3.95	6.93
$[-2,2]$	3.90	5.37	8.71
$[-3,3]$	5.34	6.87	10.46
$[-4,4]$	6.38	8.32	11.91

TABLE 1. Simulated asymptotic critical values for the asymptotic null distribution for various sets  $M$ .

### 3. NEYMAN $C(\alpha)$ TESTS FOR MIXTURE MODELS

Neyman's  $C(\alpha)$  tests can be viewed as an expanded class of Rao (score) tests that accommodate general methods of estimation for nuisance parameters. In regular likelihood settings  $C(\alpha)$  tests are constructed from the usual score components. Suppose we have iid  $X_1, X_2, \dots, X_n$  from the density  $\varphi(x, \vartheta, \xi)$ , and we would like to test the hypothesis,  $H_0 : \xi = \xi_0$  on a  $p$  dimensional parameter versus the alternative  $H_1 : \xi \neq \xi_0$ . Given a  $\sqrt{n}$ -consistent estimator,  $\hat{\vartheta}_n$ , of the nuisance parameter  $\vartheta$ , we will denote by

$$C_{\xi,n} = n^{-1/2} \sum_{i=1}^n \nabla_{\xi} \log \varphi(X_i, \vartheta, \xi) \Big|_{\substack{\xi=\xi_0 \\ \vartheta=\hat{\vartheta}_n}}$$

$$C_{\vartheta,n} = n^{-1/2} \sum_{i=1}^n \nabla_{\vartheta} \log \varphi(X_i, \vartheta, \xi) \Big|_{\substack{\xi=\xi_0 \\ \vartheta=\hat{\vartheta}_n}}$$

the score vectors with respect to  $\xi$  and  $\vartheta$  respectively. Following Akritas (1988) and Chibisov (1973) the  $C(\alpha)$  test of  $H_0$  can be viewed (asymptotically) as a conditional test in the limiting Gaussian experiment. In the limit experiment,  $(C_{\xi,n}, C_{\vartheta,n})$  are jointly Gaussian with Fisher information covariance matrix,

$$I = \begin{pmatrix} I_{\xi\xi} & I_{\vartheta\xi} \\ I_{\xi\vartheta} & I_{\vartheta\vartheta} \end{pmatrix}.$$

The conditional test of  $H_0$  based on a single observation from this limit distribution depends on the vector,

$$g_n = C_{\xi,n} - I_{\xi\vartheta} I_{\vartheta\vartheta}^{-1} C_{\vartheta,n},$$

that has covariance matrix,  $I_{\xi\xi} - I_{\xi\vartheta} I_{\vartheta\vartheta}^{-1} I_{\vartheta\xi}$ , which is the inverse of the  $\xi\xi$ -block of the inverse,  $I^{\xi\xi}$ , of the full Fisher information matrix. Thus, the  $C(\alpha)$  test statistic,  $T_n = g_n^{\top} I^{\xi\xi} g_n$ , is asymptotically  $\chi_p^2$ , and is locally optimal for alternatives of the form,  $\xi_n = \xi_0 + \delta/\sqrt{n}$ . The salient practical advantage of  $T_n$ , lies in the option to use *any*  $\sqrt{n}$ -consistent estimator for  $\hat{\vartheta}_n$ . When  $\hat{\vartheta}_n$  is the maximum likelihood estimator of  $\vartheta$  under the  $\xi = \xi_0$  constraint the  $C(\alpha)$  procedure reverts to the Rao score test.

Regularity conditions for the foregoing results were originally given by Neyman (1959) and extended by Bühler and Puri (1966) as variants of the classical Cramér conditions. An alternative formulation can be constructed from the differentiability in quadratic mean

(DQM) condition of LeCam (1970). The latter approach, as discussed in more detail in Gu (2012), seems to be more appropriate for the consideration of  $C(\alpha)$  testing for mixtures.

$C(\alpha)$  tests for heterogeneity in mixture models typically take a simple form although their theory requires some substantial amendment from the regular cases we have just described. Suppose we have random variables  $\{X_1, \dots, X_n\}$  with  $X_i \sim \varphi(x, \vartheta_i)$  and the  $\vartheta_i$ 's are given by

$$\vartheta_i = \vartheta + \tau \xi U_i$$

where the  $U_i$  are iid with distribution function,  $F$ , with  $\mathbb{E}(U) = 0$  and  $\mathbb{V}(U) = 1$ . The parameter  $\tau$  denotes a known scale parameter, and we are interested in testing the null hypothesis,  $H_0 : \xi = 0$ . Under these circumstances it can be easily seen that the usual score test procedure breaks down, because the first logarithmic derivative of the density with respect to  $\xi$  is identically zero under the null,

$$\frac{\partial}{\partial \xi} \log \int \varphi(x, \vartheta + \tau \xi u) dF(u) |_{\xi=0} = \tau \int u dF(u) \cdot \varphi'_0(x, \vartheta) / \varphi_0(x, \vartheta) = 0,$$

and consequently the usual Fisher information about  $\xi$  is zero. All is not lost, as Neyman was already aware, we can simply differentiate the log likelihood once again,

$$\begin{aligned} \frac{\partial^2}{\partial \xi^2} \log \int \varphi(x, \vartheta + \tau \xi u) dF(u) |_{\xi=0} \\ = \tau^2 \int u^2 dF(u) \cdot \varphi''_0(x, \vartheta) / \varphi_0(x, \vartheta) - \left( \tau \int u dF(u) \cdot (\varphi'_0(x, \vartheta) / \varphi_0(x, \vartheta)) \right)^2, \\ = \tau^2 \varphi''_0(x, \vartheta) / \varphi_0(x, \vartheta). \end{aligned}$$

This second-order score function replaces the familiar first-order one and provides an analogue of Fisher information for  $C(\alpha)$  parameter heterogeneity inference.

**3.1. Asymptotic Theory for  $C(\alpha)$  Tests of Parameter Heterogeneity.** The locally asymptotic normal (LAN) apparatus of LeCam can be brought to bear to establish the large sample behavior of the  $C(\alpha)$  test. We will sketch the argument in the simplest scalar parameter case, referring the reader to Gu (2012) for further details.

Let  $\{X_1, \dots, X_n\}$  be a random sample from the density  $p(x|\xi, \vartheta)$ , with respect to the measure,  $\mu$ . We would like to test the composite null hypothesis,  $H_0 : \xi = \xi_0 \in \Xi \subset \mathbb{R}$  in the presence of the nuisance parameter,  $\vartheta \in \Theta \subset \mathbb{R}^p$ , against  $H_1 : \xi \in \Xi \setminus \{\xi_0\}$ .

**Assumption 3.1.** *The density function  $p$  satisfies the following conditions:*

- (i)  $\xi_0$  is an interior point of  $\Xi$
- (ii) For all  $\vartheta \in \Theta$  and  $\xi \in \Xi$ , the density is twice continuously differentiable with respect to  $\xi$  and once continuously differentiable with respect to  $\vartheta$  for all  $x$ .
- (iii) Denoting the first two derivatives of the density with respect to  $\xi$  evaluated under the null as  $\nabla_\xi p(x|\xi_0, \vartheta)$  and  $\nabla_\xi^2 p(x|\xi_0, \vartheta)$ , we have  $\mathbb{P}(\nabla_\xi p(x|\xi_0, \vartheta) = 0) = 1$  and  $\mathbb{P}(\nabla_\xi^2 p(x|\xi_0, \vartheta) \neq 0) > 0$ .
- (iv) Denoting the derivative of the density with respect to  $\vartheta$  evaluated under the null as  $\nabla_\vartheta p(x|\xi_0, \vartheta)$ , for any  $p$ -dimensional vector  $a$ ,  $\mathbb{P}(\nabla_\xi^2 p(x|\xi_0, \vartheta) \neq a^\top \nabla_\vartheta p(x|\xi_0, \vartheta)) > 0$ .



The crucial additional requirement is that  $p$  satisfies the following modified version of LeCam's differentiability in quadratic mean (DQM) condition. The LeCam approach has two salient advantages under the present circumstances: it avoids making superfluous further differentiability assumptions, and it removes any need for the symmetry assumption on the distribution of the heterogeneity that frequently appears in earlier examples of such tests.

**Definition 3.2.** *The density  $p(x|\xi, \vartheta)$  satisfies the modified differentiability in quadratic mean condition at  $(\xi_0, \vartheta)$  if there exists a vector  $v(x) = (v_1(x), v_2(x)) \in L_2(\mu)$  such that as  $(\xi_n, \vartheta_n) \rightarrow (\xi_0, \vartheta)$ ,*

$$\int |\sqrt{p(x|\xi_n, \vartheta_n)} - \sqrt{p(x|\xi_0, \vartheta)} - h_n^\top v(x)|^2 \mu(dx) = o(\|h_n\|^2)$$

where  $h_n = ((\xi_n - \xi_0)^2, (\vartheta_n - \vartheta)^\top)^\top$ . Let  $\beta(h_n)$  be the mass of the part of  $p(x|\xi_n, \vartheta_n)$  that is  $p(x|\xi_0, \vartheta)$ -singular, then as  $(\xi_n, \vartheta_n) \rightarrow (\xi_0, \vartheta)$ ,  $\beta(h_n)/\|h_n\|^2 \rightarrow 0$ .

The second-order score function with respect to  $\xi$  implies that the corresponding term in  $h_n$  is quadratic, and this in turn implies the  $O(n^{-1/4})$  rate for the local alternative in the following theorem.

**Theorem 3.3.** *Suppose  $(X_1, \dots, X_n)$  are iid with density  $p$  satisfying Assumption 3.1 and the modified DQM condition with,*

$$v(x) = (v_1(x), v_2^\top(x))^\top = \left( \frac{1}{4} \frac{\nabla_\xi^2 p(x|\xi_0, \vartheta)}{\sqrt{p(x|\xi_0, \vartheta)}} \mathbb{I}_{[p(x|\xi_0, \vartheta) > 0]}, \frac{1}{2} \frac{\nabla_\vartheta p(x|\xi_0, \vartheta)^\top}{\sqrt{p(x|\xi_0, \vartheta)}} \mathbb{I}_{[p(x|\xi_0, \vartheta) > 0]} \right)^\top,$$

Denote the joint distribution of the  $X_i$ 's by  $P_{n, \xi, \vartheta}$ . Then for fixed  $\delta_1$  and  $\delta_2$ , the log-likelihood ratio has the following quadratic approximation under the null:

$$\Lambda_n = \log \frac{dP_{n, \xi_0 + \delta_1 n^{-1/4}, \vartheta + \delta_2 n^{-1/2}}}{dP_{n, \xi_0, \vartheta}} = t^\top S_n - \frac{1}{2} t^\top J t + o_P(1)$$

where  $t = (\delta_1^2, \delta_2^\top)^\top$ ,

$$S_n = \begin{pmatrix} S_{1n} \\ S_{2n} \end{pmatrix} = \begin{pmatrix} \frac{2}{\sqrt{n}} \sum_i \frac{v_1(x_i)}{\sqrt{p(x_i|\xi_0, \vartheta)}} \\ \frac{2}{\sqrt{n}} \sum_i \frac{v_2(x_i)}{\sqrt{p(x_i|\xi_0, \vartheta)}} \end{pmatrix}$$

and

$$J = 4 \int (vv^\top) \mu(dx) = \begin{pmatrix} \mathbb{E}(S_{1n}^2) & \text{cov}(S_{1n}, S_{2n}^\top) \\ \text{cov}(S_{1n}, S_{2n}) & \mathbb{E}(S_{2n} S_{2n}^\top) \end{pmatrix} \equiv \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix}.$$

Given that the conditions for the log likelihood ratio expansion are met local asymptotic optimality and the distribution of the  $C(\alpha)$  test statistic under appropriate local alternatives follows.

**Theorem 3.4.** *Let  $\varepsilon_n$  be a sequence of experiments based on iid random variables  $(X_1, \dots, X_n)$  with joint distribution  $P_{n, \xi_0 + \delta_1 n^{-1/4}, \vartheta + \delta_2 n^{-1/2}}$  indexed by  $t = (\delta_1^2, \delta_2^\top)^\top \in \mathbb{R}_+ \times \mathbb{R}^p$ , such that the log-likelihood ratio satisfies,*

$$\log \left( \frac{dP_{n, \xi_0 + \delta_1 n^{-1/4}, \vartheta + \delta_2 n^{-1/2}}}{dP_{n, \xi_0, \vartheta}} \right) = t^\top S_n - \frac{1}{2} t^\top J t + o_P(1),$$

*with the sequence  $S_n$  converging in distribution under the null to  $\mathcal{N}(0, J)$ . Then the sequence  $\varepsilon_n$  converges to the limit experiment based on observing one sample from  $Y = t + v$  where  $v \sim \mathcal{N}(0, J^{-1})$ . The locally asymptotically optimal statistic for testing,  $H_0 : \delta_1 = 0$  vs.  $H_1 : \delta_1 \neq 0$  is*

$$Z_n = (J_{11} - J_{12} J_{22}^{-1} J_{21})^{-1/2} (S_{1n} - J_{12} J_{22}^{-1} S_{2n}).$$

*It has distribution  $\mathcal{N}(0, 1)$  under  $H_0$ , and distribution  $\mathcal{N}(\delta_1^2 (J_{11} - J_{12} J_{22}^{-1} J_{21})^{1/2}, 1)$  under  $H_1$ . We reject  $H_0$  if  $(0 \vee Z_n)^2 > c_\alpha$  with  $c_\alpha$ , the  $(1-\alpha)$  quantile of the  $\frac{1}{2} \chi_0^2 + \frac{1}{2} \chi_1^2$  distribution.*

**Remark 3.5.** The behavior of the test statistic under the specified contiguous alternatives, follows from LeCam's third lemma. Under the null, we have

$$(Z_n, \Lambda_n) \overset{P_{n, \xi_0, \vartheta}}{\rightsquigarrow} \mathcal{N} \left( \begin{pmatrix} 0 \\ -\frac{1}{2} t^\top J t \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & t^\top J t \end{pmatrix} \right)$$

with  $\sigma_{12} = \text{cov}(Z_n, \Lambda_n) = \delta_1^2 (J_{11} - J_{12} J_{22}^{-1} J_{21})^{1/2}$ . LeCam's third lemma then implies that under the local alternative with  $\xi_n = \xi_0 + \delta_1 n^{-1/4}$  and  $\vartheta_n = \vartheta + \delta_2 n^{-1/2}$ ,

$$Z_n \overset{P_{n, \xi_n, \vartheta}}{\rightsquigarrow} \mathcal{N}(\sigma_{12}, 1).$$

Note that the test statistic  $Z_n$  is a function of  $\vartheta$ . All of the results above hold if the true nuisance parameter  $\vartheta$  is used in the test statistic, which is infeasible. In practise, we can plug in a consistent estimator for  $\vartheta$ , say  $\hat{\vartheta}$ . In order for the preceding results to be useful, we need to ensure that  $Z_n(\hat{\vartheta}) - Z_n(\vartheta) = o_P(1)$ . There are various ways to obtain this kind of result. The classical approach by Neyman (1959) was to make additional differentiability and boundedness assumptions on the function  $g$ , which is defined as

$$g(x_i, \vartheta) = (J_{11} - J_{12} J_{22}^{-1} J_{21})^{-1/2} \left( \frac{2v_1(x_i)}{\sqrt{p(x_i; \xi_0, \vartheta)}} - J_{12} J_{22}^{-1} \frac{2v_2(x_i)}{\sqrt{p(x_i; \xi_0, \vartheta)}} \right)$$

such that  $Z_n(\vartheta) = \frac{1}{\sqrt{n}} \sum_i g(x_i, \vartheta)$ . Details of these assumptions can be found in Neyman (1959, Definition 3). The assumptions are rather strong since they require the density to be three time differentiable with respect to  $\vartheta$ . Another approach, however, is to view the difference  $Z_n(\vartheta) - Z_n(\hat{\vartheta})$  as an empirical process. More precisely, if the following condition holds, we can show the difference goes to zero in probability. Details can be found in Gu (2012).

**Assumption 3.6.** *Assume that for every  $\vartheta \in \Theta$  there exists some  $\delta > 0$  such that for any  $\eta, \eta' \in U_\delta(\vartheta)$  we have for some  $\gamma > 0$*

$$|g(x, \eta) - g(x, \eta')| \leq \|\eta - \eta'\|^\gamma H(x)$$

for  $P_{n,\xi_n,\vartheta}$ -almost all  $x$  (for every  $n \in \mathbb{N}$ ) where  $H$  is square integrable with respect to  $P_{n,\xi_n,\vartheta}$  for all  $n \in \mathbb{N}$ ,  $\sup_n \mathbb{E}_{P_{n,\xi_n,\vartheta}} H^2(X) < \infty$  and additionally for some  $c_n = o(1)$   $n^{1/2} \mathbb{E}_{P_{n,\xi_n,\vartheta}} [H(X) I\{H(X) > n^{1/2} c_n\}] = o(1)$ .

**Theorem 3.7.** *Under Assumptions 3.1, 3.6 and the DQM condition, if  $\hat{\vartheta}$  is a consistent estimator for  $\vartheta$ , then*

$$|Z_n(\hat{\vartheta}) - Z_n(\vartheta)| = o_P(1)$$

**Example 3.8.** Consider testing for homogeneity in the Gaussian location mixture model with independent observations  $X_i \sim \mathcal{N}(\vartheta_i, 1)$ ,  $i = 1, \dots, n$ . Assume that  $\vartheta_i = \vartheta_0 + \tau \xi U_i$ , for known  $\tau$ , and iid  $U_i \sim F$  with  $\mathbb{E}U = 0$  and  $\mathbb{V}U = 1$ . We would like to test  $H_0 : \xi = 0$  with the location parameter  $\vartheta_0$  treated as a nuisance parameter. The second-order score for  $\xi$  is found to be,  $\nabla_\xi^2 \log \varphi(x, \vartheta_0, \xi = 0) = \tau^2((x - \vartheta_0)^2 - 1)$  and the first-order score for  $\vartheta_0$  is,  $\nabla_{\vartheta_0} \log \phi(x, \vartheta_0, \xi = 0) = (x - \vartheta_0)$ . Note that under the null,  $J_{12} = \text{cov}(\nabla_\xi^2 \log \varphi(X, 0, \vartheta_0), \nabla_{\vartheta_0} \log \varphi(X, 0, \vartheta_0)) = 0$ . Thus, we have the locally asymptotically optimal  $C(\alpha)$  test as

$$Z_n = \frac{1}{\sqrt{2n}} \sum_{i=1}^n ((X_i - \vartheta_0)^2 - 1)$$

The obvious estimate for the nuisance parameter is the sample mean, and we reject the null hypothesis when  $(0 \vee Z_n)^2 > c_\alpha$ .

#### 4. SOME SIMULATION EVIDENCE

To explore the finite sample performance of the methods we have discussed we begin with an experiment to compare the critical values of the LRT of homogeneity in the Gaussian location model with the simulated asymptotic critical values of Table 1. We consider sample sizes,  $n \in \{100, 500, 1000, 5000, 10000\}$  and four choices of the domain of the MLE of the mixture is estimated:  $\{[-j, j] : j = 1, \dots, 4\}$ . We maintain a grid spacing of 0.01 for the mixing distribution on these domains for each of these cases for the Kiefer-Wolfowitz MLE. Results are reported in Table 2. For the three largest sample sizes we bin the observations into 300 and 500 equally spaced bins respectively. It will be noted that the empirical critical values are consistently smaller than those simulated from the asymptotic theory. There appears to be a tendency for the empirical critical values to increase with  $n$ , but this tendency is rather weak. This finding is perhaps not entirely surprising in view of the slow rates of convergence established elsewhere in the literature, see e.g. Bickel and Chernoff (1993) and Hall and Stewart (2005).

To compare power of the  $C(\alpha)$  and LRT to detect heterogeneity in the Gaussian location model we conducted four distinct experiments. Two were based on variants of the Chen (1995) example with the discrete mixing distribution  $F(\vartheta) = (1 - \lambda)\delta_{h/(1-\lambda)} + \lambda\delta_{-h/\lambda}$ . In the first experiment we set  $\lambda = 1/3$ , as in the original Chen example, in the second experiment we set  $\lambda = 1/20$ . We consider four tests: (i) the  $C(\alpha)$  as described in Example 3.8, (ii.) a parametric version of the LRT in which only the value of  $h$  is assumed to be unknown and

n	cval(.90)				cval(.95)				cval(.99)			
	[-1,1]	[-2,2]	[-3,3]	[-4,4]	[-1,1]	[-2,2]	[-3,3]	[-4,4]	[-1,1]	[-2,2]	[-3,3]	[-4,4]
100	2.09	2.69	2.80	2.80	3.07	3.70	3.97	4.06	6.43	7.58	8.31	8.55
500	2.22	2.80	2.96	2.98	3.06	3.87	4.41	4.41	5.69	7.07	7.45	7.52
1,000	2.67	3.46	3.72	3.76	3.73	4.95	5.44	5.56	7.26	8.55	9.51	9.76
5,000	2.68	3.56	3.91	3.96	3.79	4.54	4.83	5.09	6.52	8.15	8.32	8.38
10,000	2.41	3.11	3.29	3.46	3.61	4.45	4.72	4.97	6.23	7.51	7.96	8.32
$\infty$	2.75	3.90	5.34	6.38	3.95	5.37	6.87	8.32	6.93	8.71	10.46	11.91

TABLE 2. Critical Values for Likelihood Ratio Test of Gaussian Parameter Homogeneity: The first five rows of the table report empirical critical values based on 1000 replications of the LRT based on the Kiefer-Wolfowitz estimate of the nonparametric Gaussian location mixture distribution. Results for sample sizes 5,000 and 10,000 were computed by binning the observations into 300, 500 equally spaced bins respectively. Restriction of the domain of the mixing distribution is indicated by the column labels. The last row reproduces the simulated asymptotic critical values reported in Table 1.

the relative probabilities associated with the two mass points are known; this enables us to relatively easily find the MLE,  $\hat{h}$  by separately optimizing the likelihood on the positive and negative half-line and taking the best of the two solutions, (iii.) the Kiefer-Wolfowitz LRT computed with equally spaced binning on the support of the sample, and finally as benchmark (iv.) the classical Kolmogorov-Smirnov test of normality. The sample size in all the power comparisons was taken to be 200, with 10,000 replications. We consider 21 distinct values of  $h$  for each of the experiments equally spaced on the respective plotting regions.

In the left panel of Figure 1 we illustrate the results for the first experiment with  $\lambda = 1/3$ :  $C(\alpha)$  and the parametric LRT are essentially indistinguishable in this experiment, and both have slightly better performance than the nonparametric LRT. All three of these tests perform substantially better than the Kolmogorov-Smirnov test. In the right panel of Figure 1 we have results of another version of the Chen example, except that now  $\lambda = 1/20$ , so the mixing distribution is much more skewed. Still  $C(\alpha)$  does well for small values of  $h$ , but for  $h \geq 0.07$  the two LRT procedures, which are now essentially indistinguishable, dominate. Again, the KS test performance is poor compared to the other tests explicitly designed for the mixture setting.

In Figure 2 we illustrate the results of two additional experiments, both of which are based on mixing distributions with densities with respect to Lebesgue measure. On the left we consider  $F(\vartheta, h) = I(-h < \vartheta < h)/(2h)$ . Again, we can reduce the parametric LRT to optimizing separately over the positive and negative half-lines to compute the MLE,  $\hat{h}$ . This would seem to give the parametric LRT a substantial advantage over the Kiefer-Wolfowitz nonparametric MLE, however as is clear from the figure there is little difference in their performance. Again, the  $C(\alpha)$  test is somewhat better than either of the LRTs, but the difference is modest. In the right panel of Figure 2 we have a similar setup, except that now the mixing distribution is Gaussian with scale parameter  $h$ , and again the ordering is

very similar to the uniform mixing case. In both of the latter experiments, the parametric LRT is somewhat undersized at the null; so we made an empirical size adjustment the two LRT curves. In all four figures the KW-LRT has been similarly size adjusted according to its performance under the null in the respective experiments.

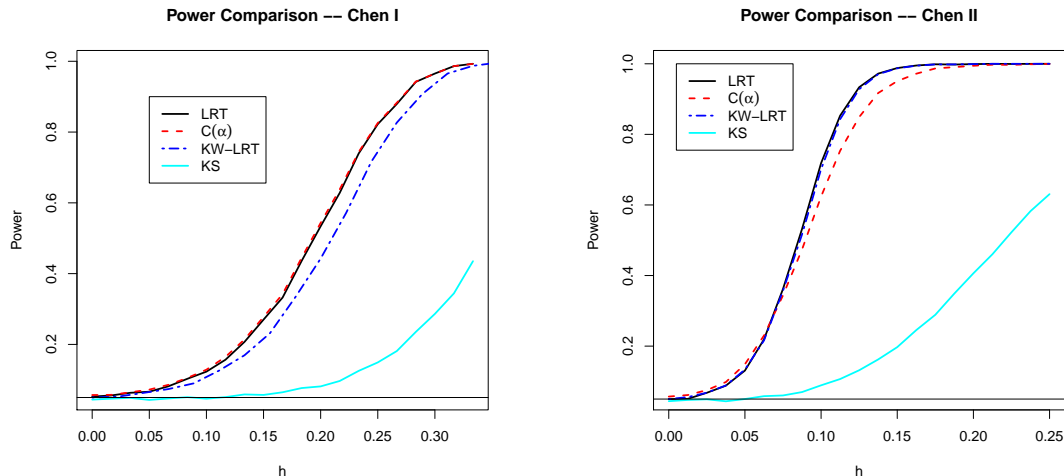


FIGURE 1. Power Comparison of Several Tests of Parameter Homogeneity: The left panel illustrates empirical power curves for four tests of parameter homogeneity for the Chen (1995) mixture with  $\lambda = 1/3$ , in the right panel we illustrate the power curves for the same four tests for the Chen mixture with  $\lambda = 1/20$ . Note that in the more extreme (right) setting, the LRTs outperform the  $C(\alpha)$  test.

## 5. CONCLUSION

We have seen that the Neyman  $C(\alpha)$  test provides a simple, powerful, albeit irregular, strategy for constructing tests of parameter homogeneity. Many examples of such tests already appear in the literature, however the LeCam apparatus provides a unified approach for studying their asymptotic behavior that enables us to relax moment conditions employed in prior work. In contrast, likelihood ratio testing for mixture models has been somewhat inhibited by their apparent computational difficulty, as well as the complexity of its asymptotic theory. Recent developments in convex optimization have dramatically reduced the computational effort of earlier EM methods, and new theoretical developments have led to practical simulation methods for large sample critical values for the Kiefer-Wolfowitz nonparametric version of the LRT. Local asymptotic optimality of the  $C(\alpha)$  test assures that it is highly competitive in most circumstances, but we have illustrated at least one case where the LRT has a slight edge. The two approaches are complementary; clearly there is little point in testing for heterogeneity if there is no mechanism for estimating models under the alternative. Since parametric mixture models are notoriously tricky to estimate,

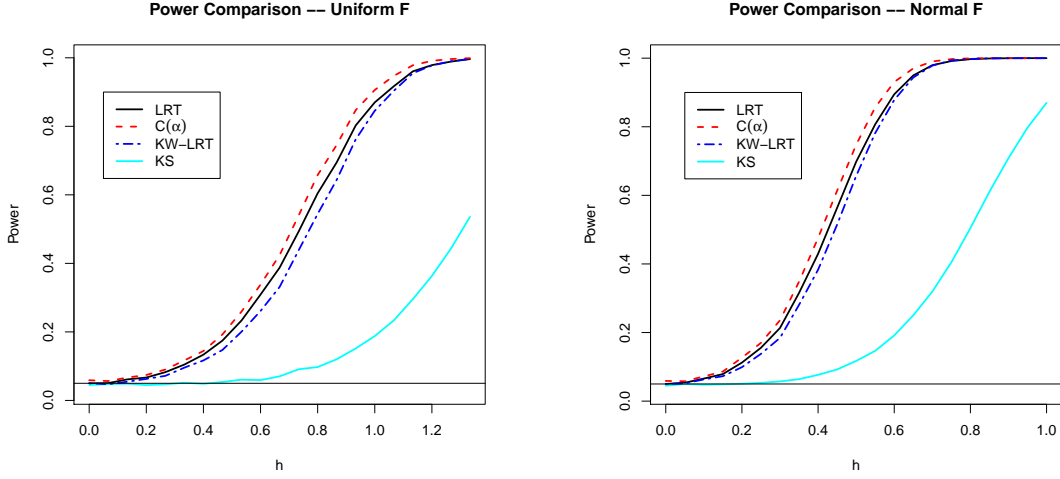


FIGURE 2. Power Comparison of Several Tests of Parameter Homogeneity: The left panel illustrates empirical power curves for four tests of parameter homogeneity for uniform mixtures of Gaussians with  $\vartheta$  on  $[-h, h]$ , on the right panel the same four power curves are depicted for Gaussian mixtures of Gaussians with standard deviation  $h$ .

it is a remarkable fact that the nonparametric formulation of the MLE problem a la Kiefer-Wolfowitz can be solved quite efficiently – even for large sample sizes by binning – and effectively used as an alternative testing procedure. We hope that these new developments will encourage others to explore these methods.

## REFERENCES

- AKRITAS, M. (1988): “An asymptotic derivation of Neyman’s  $C(\alpha)$  test,” *Statistics & Probability Letters*, 6(5), 363–367.
- ANDERSEN, E. D. (2010): “The MOSEK Optimization Tools Manual, Version 6.0,” Available from <http://www.mosek.com>.
- AZAÏS, J.-M., E. GASSIAT, AND C. MERCADIER (2009): “The likelihood ratio test for general mixture models with or without structural parameter,” *ESAIM: Probability and Statistics*, 13, 301–327.
- BICKEL, P., AND H. CHERNOFF (1993): “Asymptotic distribution of the likelihood ratio statistic in a prototypical nonregular problem,” in *Statistics and Probability: A Raghu Raj Bahadur Festschrift*, ed. by J. Ghosh, S. Mitra, K. Parthasarathy, and B. PrakasaRao, pp. 83–96. Wiley, New Delhi.
- BÜHLER, W., AND P. PURI (1966): “On optimal asymptotic tests of composite hypotheses with several constraints,” *Probability Theory and Related Fields*, 5, 71–88.
- CHEN, J. (1995): “Optimal rate of convergence for finite mixture models,” *The Annals of Statistics*, 23, 221–233.
- CHIBISOV, D. M. (1973): “Asymptotic expansions for Neyman’s  $C(\alpha)$  tests,” in *Proceedings of the Second Japan-USSR Symposium on Probability Theory, Lecture Notes in Mathematics*, vol. 330, pp. 16–45. Institute for Mathematical Statistics.
- FRIBERG, H. A. (2012): *Rmosek: The R-to-MOSEK Optimization Interface*, R package version 1.2.3.
- GU, J. (2012): “Neyman’s  $C(\alpha)$  Test for Unobserved Heterogeneity,” preprint.

- HALL, P., AND M. STEWART (2005): “Theoretical analysis of power in a two-component normal mixture model,” *Journal of statistical planning and inference*, 134, 158–179.
- HECKMAN, J., AND B. SINGER (1984): “A method for minimizing the impact of distributional assumptions in econometric models for duration data,” *Econometrica*, 52, 63–132.
- JIANG, W., AND C.-H. ZHANG (2009): “General maximum likelihood empirical Bayes estimation of normal means,” *Annals of Statistics*, 37, 1647–1684.
- KIEFER, J., AND J. WOLFOWITZ (1956): “Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters,” *The Annals of Mathematical Statistics*, 27, 887–906.
- KOENKER, R. (2012): *REBayes: Empirical Bayes Estimation and Inference in R*, R package version 0.23.
- KOENKER, R., AND I. MIZERA (2012): “Convex Optimization, Shape Constraints, Compound Decisions and Empirical Bayes Rules,” preprint.
- LAIRD, N. (1978): “Nonparametric Maximum Likelihood Estimation of a Mixing Distribution,” *Journal of the American Statistical Association*, 73, 805–811.
- LECAM, L. (1970): “On the assumptions used to prove asymptotic normal of maximum likelihood estimators,” *The Annals of Mathematical Statistics*, pp. 802–828.
- LINDSAY, B. (1983): “The Geometry of Mixture Likelihoods: A General Theory,” *Annals of Statistics*, 11, 86–94.
- (1995): *Mixture Models: Theory, Geometry and Applications*. NSF-CBMS-IMS Conference Series in Statistics, Hayward, CA.
- LIU, X., AND Y. SHAO (2003): “Asymptotics for likelihood ratio tests under loss of identifiability,” *The Annals of Statistics*, 31, 807–832.
- NEYMAN, J. (1959): “Optimal Asymptotic Tests of Composite Statistical Hypotheses,” in *Probability and Statistics, the Harald Cramer Volume*, ed. by U. Grenander. Wiley: New York.
- ROBBINS, H. (1950): “A Generalization of the Method of Maximum Likelihood: Estimating a Mixing Distribution (Abstract),” *The Annals of Mathematical Statistics*, 21, 314.
- SAUNDERS, M. A. (2003): “PDCO: A Primal-Dual interior solver for convex optimization,” <http://www.stanford.edu/group/SOL/software/pdco.html>.

#### APPENDIX A. TECHNICAL DETAILS

**Proof of (2)** Given a measure  $G \in P_M, G \neq \delta_0$  define  $V(G) := \sum_{k=2}^{\infty} \frac{\kappa_k^2(G)}{k!}$ . Also, define for  $n \in \mathbb{N}$  and  $\alpha \in [-N, N]$  the probability measure  $\tilde{G}_n := p_n \delta_{c_n} + (1 - p_n)G$  with  $p_n := 1 - V(G)/n$  and  $c_n := \frac{1-p_n}{p_n}(\alpha - \kappa_1(G))$  [the dependence of  $p_n, c_n$  on  $G$  is suppressed in the notation]. Note that for  $n$  sufficiently large we have  $\tilde{G}_n \in P_M$  for all  $\alpha \in [-N, N]$ . Moreover, by construction  $\kappa_1(\tilde{G}_n) = \alpha(1 - p_n)$  and

$$\kappa_k(\tilde{G}_n) = \kappa_k(G)(1 - p_n) + (1 - p_n) \left( \frac{1 - p_n}{p_n} \right)^{k-1} (\alpha - \kappa_1(G))^k$$

for  $n \in \mathbb{N}$ . This implies for  $n$  sufficiently large we have a.s.

$$\left| \alpha Y_1 + \sum_{k=2}^{\infty} \frac{Y_k \kappa_k(G)}{(k!)^{1/2}} - \frac{1}{1 - p_n} \sum_{k=1}^{\infty} \frac{Y_k \kappa_k(\tilde{G}_n)}{(k!)^{1/2}} \right| \leq \frac{1 - p_n}{p_n} \sum_{k=2}^{\infty} \frac{|Y_k| \tilde{C}^k}{\sqrt{k!}} \left( \frac{1 - p_n}{p_n} \right)^{k-2} \leq \frac{2\tilde{C}^2 V(G)}{n} \sum_{k=2}^{\infty} \frac{|Y_k|}{\sqrt{k!}}$$

and

$$\left| \alpha^2 + \sum_{k=2}^{\infty} \frac{\kappa_k^2(G)}{k!} - \frac{1}{(1 - p_n)^2} \sum_{k=1}^{\infty} \frac{\kappa_k^2(\tilde{G}_n)}{k!} \right| \leq \frac{CV(G)}{n}$$

for finite constants  $C, \tilde{C}$  depending only on  $N$  but not on  $\alpha$  and  $G$  [note that  $G \in P_M$  has support contained in  $[L, U]$ ]. Thus for every  $N < \infty, \varepsilon > 0$  we have with probability at

least  $1 - \varepsilon$  [this follows by choosing  $n$  sufficiently large]

$$\sup_{G \in P_M} \frac{\sum_{k=1}^{\infty} \frac{Y_k \kappa_k(G)}{(k!)^{1/2}}}{\left( \sum_{k=1}^{\infty} \frac{\kappa_k^2(G)}{k!} \right)^{1/2}} \geq \sup_{\alpha \in [-N, N]} \sup_{G \in P_M} \frac{\alpha Y_1 + \sum_{k=2}^{\infty} \frac{Y_k \kappa_k(G)}{(k!)^{1/2}}}{\left( \alpha^2 + \sum_{k=2}^{\infty} \frac{\kappa_k^2(G)}{k!} \right)^{1/2}} - \varepsilon.$$

Next, observe that  $N$  can be chosen so large that with probability at least  $1 - f(\varepsilon)$

$$\sup_{\alpha \in \mathbb{R} \setminus [-N, N]} \sup_{G \in P_M} \frac{\alpha Y_1 + \sum_{k=2}^{\infty} \frac{Y_k \kappa_k(G)}{(k!)^{1/2}}}{\left( \alpha^2 + \sum_{k=2}^{\infty} \frac{\kappa_k^2(G)}{k!} \right)^{1/2}} \leq |Y_1| + \varepsilon$$

where  $f(a) \rightarrow 0$  for  $a \rightarrow 0$ . Finally, note that

$$\sup_{G \in P_M} \frac{\sum_{k=1}^{\infty} \frac{Y_k \kappa_k(G)}{(k!)^{1/2}}}{\left( \sum_{k=1}^{\infty} \frac{\kappa_k^2(G)}{k!} \right)^{1/2}} \geq |Y_1| \quad \text{a.s.}$$

[consider the sequence of measures  $G_n = \delta_{\text{sign}(Y_1)/n} \in P_M$ ].

Summarizing the findings above, we have shown that for any  $\varepsilon > 0$  we have with probability arbitrarily close to one:

$$\sup_{G \in P_M} \frac{\sum_{k=1}^{\infty} \frac{Y_k \kappa_k(G)}{(k!)^{1/2}}}{\left( \sum_{k=1}^{\infty} \frac{\kappa_k^2(G)}{k!} \right)^{1/2}} \geq \sup_{\alpha \in \mathbb{R}} \sup_{G \in P_M} \frac{\alpha Y_1 + \sum_{k=2}^{\infty} \frac{Y_k \kappa_k(G)}{(k!)^{1/2}}}{\left( \alpha^2 + \sum_{k=2}^{\infty} \frac{\kappa_k^2(G)}{k!} \right)^{1/2}} - \varepsilon.$$

By letting  $\varepsilon \rightarrow 0$  the above can be turned in an almost sure inequality with no  $\varepsilon$  on the right-hand side. Finally, setting  $\alpha = \kappa_1(G)$  we see that the converse inequality also holds almost surely. Thus we have shown that

$$\sup_{G \in P_M} \frac{\sum_{k=1}^{\infty} \frac{Y_k \kappa_k(G)}{(k!)^{1/2}}}{\left( \sum_{k=1}^{\infty} \frac{\kappa_k^2(G)}{k!} \right)^{1/2}} = \sup_{\alpha \in \mathbb{R}} \sup_{G \in P_M} \frac{\alpha Y_1 + \sum_{k=2}^{\infty} \frac{Y_k \kappa_k(G)}{(k!)^{1/2}}}{\left( \alpha^2 + \sum_{k=2}^{\infty} \frac{\kappa_k^2(G)}{k!} \right)^{1/2}} \quad \text{a.s.}$$

Define  $\beta_k := \frac{\kappa_k(G)}{(k!)^{1/2}}$ . Fix a realization of  $Y_1, Y_2, \dots$ . First, observe that it suffices to consider the supremum over  $G \in P_M$  with  $\sum_{k=2}^{\infty} Y_k \beta_k \geq 0$ . Fixing  $G \in P_M$  shows that in the case  $\sum_{k=2}^{\infty} Y_k \beta_k > 0$  the supremum with respect to  $\alpha$  on the right-hand side above is attained for  $\alpha^* = Y_1 \frac{\sum_{k=2}^{\infty} \beta_k^2}{\sum_{k=2}^{\infty} Y_k \beta_k}$ , and plugging this into the equation above we obtain (after some simple algebra)

$$\sup_{\alpha \in \mathbb{R}} \frac{\alpha Y_1 + \sum_{k=2}^{\infty} \frac{Y_k \kappa_k(G)}{(k!)^{1/2}}}{\left( \alpha^2 + \sum_{k=2}^{\infty} \frac{\kappa_k^2(G)}{k!} \right)^{1/2}} = \left( Y_1^2 + \frac{\left( \sum_{k=2}^{\infty} Y_k \beta_k \right)^2}{\sum_{k=2}^{\infty} \beta_k^2} \right)^{1/2}$$

for every  $G \in P_M$  with  $\sum_{k=2}^{\infty} Y_k \beta_k > 0$ . In the case  $\sum_{k=2}^{\infty} Y_k \beta_k = 0$  we obtain  $\alpha^* = \text{sign}(Y_1)$ . Summarizing the above arguments yields

$$\left( \sup_{G \in P_M} \left( \frac{\sum_{k=1}^{\infty} \frac{Y_k \kappa_k(G)}{(k!)^{1/2}}}{\left( \sum_{k=1}^{\infty} \frac{\kappa_k^2(G)}{k!} \right)^{1/2}} \right)_+ \right)^2 = Y_1^2 + \sup_{G \in P_M} \frac{\left( \left( \sum_{k=2}^{\infty} Y_k \beta_k \right)_+ \right)^2}{\sum_{k=2}^{\infty} \beta_k^2}$$



and this directly implies (2)

□